关联开放数据在图书情报服务的应用逻辑分析*

顾立平
(中国科学院国家科学图书馆 北京 100190)

【摘要】开放关联数据是开放科学中的开放数据集成应用的一项重要部分,简介关联开放数据的应用需求与范畴;聚焦图书馆应用开放关联数据的实施方案,包括如何转化书目数据与使用数据为开放数据,以及如何应用社会网络和社会标签这类开放数据;系统性分析书目品质提升、电子资源管理、信息计量分析、专利竞争情报、数字典藏、嵌入式学科咨询等的关联开放数据的技术路线。

【关键词】开放科学 开放数据 关联数据 语义网 本体 社会网络 社会标签 社会媒体 社群空间
【分类号】G250

Analysis for Logic of the Linked Open Data in the Library and Information Services

Ku Liping
(National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

【Abstract】The Linked Open Data (LOD) is one important part of the open data application and integration under the open science. This article introduces briefly the application need and field of the LOD, focuses on the implementation of the library using LOD include how to transfer the bibliography and the usage data into open data and how to use the open data as like social network and social tagging. The author also systemically analyses the technology roadmap of the improvement by using LOD on bibliographic quality, electronic resource management, informatics application, patent competitive intelligence and digital archives.

【Keywords】Open science Open data Linked data Semantic Web Ontology Social network Social tag Social media Social community space

1 图书馆在开放科学时代中处理开放数据的抉择

重复使用刊载在科学出版物上的知识以及公开批判,是科学的两大基础;开放科学数据是促使科研工作有效运作以及社会充分从科学劳动中获益的关键,其基础是有效地传播具有开放性(Openness)的科研数据[1,2]。在网络时代,能够支持开放知识、开放内容与开放服务的开放性具有 11 项特征[3]:近用/获取(Access)、重复散布、重复使用、无技术限制、署名、完整、平等对待用户、平等看待领域、授权散布、授权条款不得专属于特定产品、授权条款不得限制散布其他作品等。因此,并非所有网络信息都具备开放性的特征,正因如此,充分推展以及利用现有

收稿日期: 2012-06-12
收修改稿日期:2012-09-13
*本文系中国科学院国家科学图书馆指向性人才项目"科技信息政策研究与咨询中心"(项目编号:馆 1203)的研究成果之一。

的开放数据,成为当代科学发展的关键问题。

另一方面,网络信息资源具有异类(不同内容类型和体裁)、异构(不同的数据格式及相应的语义规则)和分布式的特点,不利于将信息资源集中一处(如图书馆)的传统处理方式;作为一种新兴的知识组织和知识发现方式,关联数据(Linked Data)提供了一套低成本的标准化数据访问机制,并且可以有效地规避一些复杂的数据权益纷争问题;然而,在发布现有资源的Linked Data过程中,存在因为不同服务方式、不同分类体系、不同数据内容,以及不同领域中采用不同的本体及词表所带来的复杂性[4]。这项复杂性又带来交互界面、关系有效性、数据融合映射、数据许可权等问题[5]。作为语义网发展的重要推力之一,关联数据通过网络把以前没有关联的相关数据连接起来,其关键是[6]采集数据、公布关联数据、规范化链接中的“连接点”、从个别数据集获得数据以及客户端应用程序数据等。处理程序和规则的关联开放数据(Linked Open Data,LOD)包括[6]:标准(Standards)、获取(Access)、许可(License)、识别符 Identifiers(URIs)、数据模型 Data Model(如资源描述框架 RDF)、本体(RDFS, OWL)、查询语言(SPARQL)等6个核心要素。

在网络化、数字化、开放化的科研时代中,图书情报事业必须正视开放数据及其开放科学的新兴需求,而LOD正当其时。

2 关联开放数据的形式逻辑

传统上,图书馆处理数据的方式分为数据和元数据,数据被采集和保存,狭义可指计算机中的一组位元,广义可指图书、期刊、报章、单件电子档案等内容载体,而元数据则是用以描述广义数据下的数据,如书目、索摘、分类、目录等。在此基础之上,专家可以进行获取、分析、呈现等情报工作。

理论上,快速发展的网络环境以及运算技术使数据可以自由流通、元数据可以自由创建,并且可以凭借数据集的重新组合、添加、删减等动作创建新的数据集;可以建立元数据进行元数据与元数据之间的关系,从这些关系中找到原先不被找到的数据。若以函数表示,如下:

(1)数据集 D{d1,d2,…,dn};描述 D 的元数据 M{m1d1,m2d2,…,mndn}。

(2)数据集 D’{d’1,d’2,…,d’n};描述 D 的元数据 M{m1d’1,m2d’2,…,mnd’n}。

(3)数据集 D”{d”1,d”2,…,d”n};描述 D 的元数据 M’{m’1d”1,m’2d”2,…,m’nd”n}。

根据传统方式:

(4)从 M 中提取 m1 和 m3 可得 D{d1,d3}以及 D’{d’1,d’3}。

然而,如果定义 M = M’则:

(5)从 M 中提取 m1 和 m3 可得 D{d1,d3}以及 D’{d’1,d’3}以及 D”{d”1,d”3}。

如果定义 M ~ M’其中 m1 = m’1 且 m3 ≠ m’3 则:

(6)从 M 中提取 m1 和 m3 可得 D{d1,d3}以及 D’{d’1,d’3}以及 D”{d”1}。

综上,比较(4)、(5)、(6)后,可见此策略具有两个作用:

①从被定义的新关系中找到原先不会被找到的数据;

②调整定义可以遮蔽部分数据。如此,一开一关,形成阈值。

承上,若以函数 L 表示各个 D 的访问许可,则从 M 到 D 的提取过程必需增加 L 函数,现实中的许可证的形式有许多种,如开放性的11种特征等,为求理解而简化许可证 L 仅有两种形式:

L = 0 //不允许访问

L = 1 //允许访问

由此考虑到:如果数据不开放,如何进行关联?

3 关联开放数据的三种效果

将以上疑问重设为可观察的研究问题,则是:在何种程度的开放性下能够利用关联阈值产生何种效果?如果 LOD 能够最大化数据的开放性,则有三个显著效果:

(1)对网络上非结构化文本进行结构化

虽然结构化数据量不断在网络上增加(特别是为了促进联系的主动公开数据),网络上还包括(而且主要是)非结构化数据,特别是文本内容。因此,一个重要的问题是:如何使人们在网络上能够有效地访问大量的非结构化信息?

设文本信息 T{D1, D2,…,Dn};若要 D 满足 LOD 形式逻辑,有两种方式:

①数据驱动(Data - driven):根据 D1{d1,d2,…,dn}和 D2{d1,d2,…,dn}中的相同数据 d 形成新的文本数据集 TD

{d1,d2,…,dn}。

②情景感知(Inductive):根据已知数据集 D'{d1,d2,…,dn}匹配文本信息 T{D1, D2,…,Dn}中的相似数据集 D。

将两种方式结合:本体、机械学习,信息检索、信息提取和文本挖掘的整合[7-9]是信息检索的热点问题。

(2)在信息结构化的基础上强化信息检索准确化

通过 LOD 增强基于关键字的搜索是强化网络导航系统的一项新作法[10]。根据 LOD 的形式逻辑,交替利用各种 M = M'和 M ~ M'逻辑,能够形成一个信息检索核心索引的分类机制,避免数据集 D、D'、D"原先面临的第(4)项情况。然而,存在不同的 URL 在不同数据集中的相同对象等数据异质性难题[11],解决办法是根据语义相似性度量,配合相关算法,使用本体和信息架构,提高针对同一对象的识别程度[12]。

LOD 提供实体类型的信息,可视为命名实体分类时的先验知识,用 LOD 的方法提取信息(命名实体的字符串、类型)来建立一个新的类型的知识基础,加入到现有的分类体系后,能够提高应用程序的性能。

(3)在信息检索准确化的基础上扩大信息范围

关联数据不但预先建立了数据对象之间的可靠关系,而且迅速发展的关联数据空间为构建高效的关联参考服务提供了强大的资源支持[13]。传统意义上的搜索是基于关键词汇的文件导向查询,而 LOD 则跨越不同的数据源,在诸如 RDF 格式和 SPARQL 查询语言标准的基础上,以机器可读的(Machine - readable)方式与类型相关的实体之间进行网络数据链接,它更类似于分布式或者联合数据库,但与之合作的数据源则是独立维护和更新他们各自的数据[14]。例如,基于实例的 myCBR 系统[15],经过模型生成、数据导入、相似模型、解释,以及可视化用户界面,利用 Symbolic 从不同实体的文本中提取信息,进行数据关联和信息推荐。

4　图书馆的 LOD 具体操作方案

图书馆使用关联数据技术能够将资源发布为关联数据、改善检索服务系统的效果、增强资源发现服务、实现数据融合与语义检索服务、跨机构数据存取和重用、促进学术研究和学术交流、实现图书馆与教学系统之间的集成等[16,17]。事实上,中国科学院国家科学图书馆已经将机构知识库(IR)中的实体关系发布为能够进行语义揭示的关联数据格式并且将 IR 中的实体数据进行语义标注[18]。

不过,满足开放科学所需的开放数据,不仅只有上述关联数据技术的应用,以下所述是如何导出图书馆的书目数据和使用数据,以及如何导入社会网络和社会标签的 LOD 具体方案。

4.1　将封闭数据(馆内)转换为开放数据(馆外)

(1)LOD 包括开放(标准、获取、许可)和关联(识别符、数据模型、本体、查询语言)两个部分[19]。因此,考虑共同协议能够让图书馆从关联数据拓展到开放关联数据。

(2)输出数据有多种方式,如使用一个系统模组、加配一套转档系统,或者进入文件系统内直接复制转出部分.dat 和.ini 文档等。在产生符合 XML 结构的文件过程中,不成比例的输出/输入容易造成数据更新所引起的计算机资源占用和正常工作停顿等问题,因此,建议采用不定期转储更新的步骤。

(3)涉及到获取、许可、识别符和数据模型的裸数据发布,需要经过 4 个步骤:加载 HTTP Server、描述发布的数据集、附加许可、登记数据集等。

(4)必需为数据选择合适的 RDF 词汇(例如书目数据选择 bio 等)以转换标识符和数据,接着映射和转换裸数据到 RDF 再写入 HTTP URIs 以描述资源,最后以 SPARQL 接口和 Pubby 框架的 HTML 方式呈现。

4.2　将开放数据(馆外)转化为馆内资源(馆内)

不断增长的社会网络数据,成为开放数据的一个重要组成部分,因此,新的分布式体系结构已经建立在语义网标识体系(如 RDF(S)/ OWL、RDFa、SPARQL 等)之上;新的范式不仅要让用户拥有自己的数据,也要通过语义网增加他们的数据,所以在诸如 DBpedia 的集中标注和检索组件的基础上,可以开发基于图形的 RDF 基础结构、独立于上下文语义关系图和外部信息源,如搜索引擎结果和社会性标签系统等的关联数据集等,作为替代网页排名技术的可扩展技术,利用 LOD 语义标记,生成查询返回近似结果的手段[20,21]。具体做法是利用 DBpedia 建模信息检索框架,然后,从这些概念中提取搜索历史特征向量,在用户查询的数据基础上,建立 LOD 模型[22]。不过,这种模型需要有维护批次导入和批次处理的能力。

终端用户通常根据不同目的在社会标签系统中标注条目、注释条目和描绘内容,例如列出照片中出现的

对象,或表达有关条目的上下文信息,或者自我引用和组织个人内容(如意见、评论)等;假设检索内容含有相当大比例的标签,那么,识别基本标签的个人意愿,可能有利于提高搜索和推荐过程的准确性[23]。在实践上,建议考虑更多范围取样和可持续运行的成本问题。

4.3 其他可能的发展

关联数据的本质是将本体和相关数据资源按照RDF格式的标准链接起来,同时要求支持HTTP URI访问和RDF查询语言SPARZL检索。按照这个原则,图书馆之间可以进行关联数据的应用,例如德国开发的R2R从词汇和数据集实例提升RDF的链接能力,将RDF和RDF进行关联,以达到LOD的目的[24]。然而,随着数据集的不断增长,由于数据发布的开放性、数据与上下文无关以及通用的资源标识等所引发对象共指问题[25]会困扰图书馆应用关联数据技术的决定。另一方面,关联数据技术能够处理图书馆知识资源的数据,但是用户使用行为的数据更多涉及开放性的标准、获取、许可原则,而这些数据是图书馆应用LOD实现开放数据的关键,并且是该馆独一无二的数据。

5 未来应用LOD的几个方向

(1)支撑书目数据品质保障

从事或者曾经从事编目性质工作的馆员,都明白建立和维护叙词表是复杂和艰巨的任务。如果利用来自某种大型已经是结构化文本的数据集的标签编辑词库,那么创建和维护叙词表的工作就更容易进行[26]。因为LOD提供了一组庞大的结构化数据,因此一旦转变为某个词库,其准确率相对较高,此外,即便是专家或者专业图书馆员不了解语义网及其技巧,也能够根据词库顺利进行叙词表的工作。

(2)支撑电子资源管理系统

使用语义网技术和关联数据的原则,可以克服从电子内容供应商被动接收资源清单或者在特定模块上操作的限制,提高资源列表管理工具的数据互操作性[27]。目前,所有图书馆对于电子资源管理的操作流程,必需是等待集成商提供资源列表,将之转化为资源清单上传后进行管理,这个时间取决于内容提供商“给单”的速度,以及图书馆员“键单”的工作速度;另一种方式是订购知识库,从知识库中“转单”到电子资源管理的资源清单中。应用LOD或许能够加快“取单”、“给单”、“转单”、“键单”等工作。

(3)支撑信息计量分析

关联开放数据提供增强数据可用性和实用性的机会,运用得当,能够全面提升各个领域科技发展。在大量已发表的文献中,大多数是不可机读的格式,如何进行信息抽取、组织数据并且提升准确性,就成为一项热点议题。例如,目前正在发展中的PatentEye原型系统,其目的就是针对专利文献中的化学反应,进行信息抽取工作,捕获反应物和产品标识的数据,加快科研工作的信息处理效率[28]。类似PatentEye的系统,着重借助关联开放数据,以达到文本挖掘和信息组织的功能。

(4)支撑专利竞争情报

世界知识产权组织中的国际专利分类,对专利搜索至关重要,因为它们通常作为一个搜索过程中的切入点。如果利用来自维基百科的科学门户网站对于一门科学的分类方法,则能够因为它们不同的分类方法而分配不同的类别,可以利用LOD建立专利分类集成和交互框架,实现允许不同的专利本体在广泛互动中的数据集整合[29]。目前,这是专利计量和专利竞争情报中的一项前沿研究,即利用关联开放数据,挖掘专业学科门户网站的专业学科分类,然后对照国际专利分类,收集相关专利并予以分析。

(5)支撑数字典藏的建设能力

传统上,数字典藏的含义是扫描、建档、备份和有限度地取用。然而,要扩大数字典藏的建设能力,除了经费和人力投入外,还可以依靠技术驱动数字典藏的创新工程。例如,依照创新理论的动因,应用于文化遗产领域的开放数据,可基于语义网技术和规范,创建一个综合性的语义知识库,通过LOD组成类似博物馆开放云的数据集成应用[30]。藉由LOD减少建档所需要投入的元数据编辑人员,将数据的质量控制问题变为较容易处理的数据品质评估方法,并且将优质数据转为开放数据,提供另一种域外服务。

(6)支撑学科服务的嵌入模式

工业生态学(IE)是一个新兴的研究领域,需要社群驱动数据的采集、加工、保管和共享,以及数据和知识的共享机制;由于涉及的技术和标准类型,在许多工业生态学家的正常工作范围之外,因此大批年轻学者在网络上对此展开讨论[31]。但是,人们很少注意到:图书馆学有大量数据管理的成功经验和失败教训,可

以少走弯路;研究型图书馆有专业的学科馆员和系统建制人才,可以作为支援。在交叉学科、新兴学科、试验学科中,存在许多意识到或者没有意识到自己需要关联开放数据服务的社群,这些人群是图书情报服务团队的潜在服务对象。

6　走向开放科学的图书馆 LOD 应用

LOD 还存在一些其他问题,例如,在网络环境中,人们常会发现一些值得怀疑、甚至自相矛盾的信息,判定这些信息的真实性需要借助数据溯源(Data Provenance);目前创建溯源信息通常采用标注或者查询求逆两种方式,并且存在多种溯源模型,然而确保科学流程的溯源信息的完全可重算能力是一项挑战[32]。在考虑 LOD 的同时,也必须考虑其他相关议题的进展,才能完整展现图书馆的信息图谱变迁的全貌,使图书馆更好把握开放数据的处理能力,以及在开放科学时代中更具积极性和创造性。本文在实作经验的基础上,根据原理基础,抛砖引玉,是以为文。

(致谢:感谢评委的评审和编辑部的校勘。)

参考文献:

[1] Panton Principles. Principles for Open Data in Science[EB/OL]. [2012-08-16]. http://pantonprinciples.org/.

[2] The Royal Society. Science as an Open Enterprise[EB/OL]. [2012-08-16]. http://royalsociety.org/policy/projects/science-public-enterprise/report/.

[3] Open Definition. Open Definition Version1.1[EL/OL]. [2012-08-16]. http://opendefinition.org/okd/.

[4] 沈志宏,张晓林. 关联数据及其应用现状综述[J]. *现代图书情报技术*, 2010 (11):1-9. (Shen Zhihong, Zhang Xiaolin. Linked Data and Its Applications:An Overview[J]. *New Technology of Library and Information Service*, 2010 (11):1-9.)

[5] 黄永文. 关联数据驱动的 Web 应用研究[J]. *图书馆杂志*, 2010, 29(7):55-59. (Huang Yongwen. Research on Linked Data-driven Web Applications[J]. *Library Journal*, 2010, 29(7):55-59.)

[6] Omitola T, Koumenides C L, Popov I O, et al. Put in Your Postcode, Out Comes the Data: A Case Study[C]. In: *Proceedings of the 7th Extended Semantic Web Conference*. 2010:318-332.

[7] Horrocks I. Ontologies and the Semantic Web[J]. *Communications of the ACM*, 2008,51(12):58-67.

[8] Krummenacher R, Norton B, Marte A. Towards Linked Open Services and Processes[C]. In: *Proceedings of the 3rd Future Internet Conference*. Berlin, Heidelberg: Springer-Verlag, 2010:68-77.

[9] Bloehdorn S, Blohm S, Cimiano P, et al. Combining Data-driven and Semantic Approaches for Text Mining[C]. In:*Proceedings of Foundations for the Web of Information and Services*. Springer, 2011:115-142.

[10] Waitelonis J, Sack H. Towards Exploratory Video Search Using Linked Data[C]. In: *Proceedings of the 11th IEEE International Symposium on Multimedia*. 2009:540-545.

[11] Noessner J, Niepert M, Meilicke C, et al. Leveraging Terminological Structure for Object Reconciliation[C]. In: *Proceedings of the 7th Extended Semantic Web Conference* (*ESWC2010*), Heraklion, Crete, Greece. Berlin, Heidelberg: Springer-Verlag, 2010:334-348.

[12] Ni Y, Zhang L, Qiu Z M, et al. Enhancing the Open-Domain Classification of Named Entity Using Linked Open Data[C]. In: *Proceedings of the 9th International Semantic Web Conference*. 2010:566-581.

[13] 刘媛媛,李春旺,黄永文. 基于 LOD 的关联参考服务构建研究[J]. *现代图书情报技术*, 2011(6):66-71. (Liu Yuanyuan, Li Chunwang, Huang Yongwen. Study on Building the Service of Relevance Reference Based on LOD[J]. *New Technology of Library and Information Service*, 2011(6):66-71.)

[14] Gorlitz O, Staab S. Federated Data Management and Query Optimization for Linked Open Data[A]. //New Directions in Web Data Management 1[M]. Berlin, Heidelberg: Springer-Verlag,2011: 109-137.

[15] Roth-Berghofer T, Adrian B, Dengel A. Case Acquisition from Text: Ontology-based Information Extraction with SCOOBIE for myCBR[C]. In: *Proceedings of the 18th International Conference on Case-Based Reasoning*. 2010:451-464.

[16] 黄永文. 关联数据在图书馆中的应用研究综述[J]. *现代图书情报技术*, 2010 (5):1-7. (Huang Yongwen. Research on Linked Data-driven Library Applications[J]. *New Technology of Library and Information Service*,2010(5):1-7.)

[17] 黄永文,岳笑,刘建华. 关联数据应用的体系框架及构建关联数据应用的建议[J]. *现代图书情报技术*, 2011 (9):7-13. (Huang Yongwen, Yue Xiao, Liu Jianhua. Architecture on Linked Data Based Applications and Some Suggestions for Building Linked Data Applications[J]. *New Technology of Library and Information Service*,2011 (9):7-13.)

[18] 王思丽,祝忠明. 利用关联数据实现机构知识库的语义扩展研究[J]. *现代图书情报技术*, 2011 (11):17-23. (Wang Sili, Zhu Zhongming. Study on the Semantic Expansion of Institutional Repository Based on Linked Data[J]. *New Technology of Library and Information Service*, 2011 (11):17-23.)

[19] Schomburg S. Publishing Aleph Data as Linked Open Data[OL]. [2011-09-22]. http://igelu.org/wp-content/uploads/2011/09/Schomburg_2011_igelu_final.pdf.

[20] Passant A, Breslin J G, Decker S. Rethinking Microblogging: Open, Distributed, Semantic[C]. In: *Proceedings of the 10th International Conference on Web Engineering*. 2010:263-277.

[21] Mirizzi R, Ragone A, Di Noia T, et al. Ranking the Linked Data: The Case of DBpedia[C]. In: *Proceedings of the 10th International Conference on Web Engineering*. 2010:337-354.

[22] Meij E, Bron M, Hollink L, et al. Mapping Queries to the Linking Open Data Cloud: A Case Study Using DBpedia[J]. *Journal of Web Semantics*, 2011, 9(4):418-433.

[23] Cantador I, Konstas I, Jose J M. Categorising Social Tags to Improve Folksonomy-based Recommendations[J]. *Journal of Web Semantics*, 2011, 9(1):1-15.

[24] 陶俊,孙坦,刘铮. 关联数据映射语言:R2R[J]. *中国图书馆学报*,2012,38(3):97-106.(Tao Jun, Sun Tan, Liu Zheng. Linked Dataset Mapping Language: R2R[J]. *Journal of Library Science in China*,2012,38(3):97-106.)

[25] 刘媛媛,李春旺. 关联数据的对象共指问题研究[J]. *情报理论与实践*, 2012, 35(2):46-51.(Liu Yuanyuan, Li Chunwang. Studies on Object Co-reference in Linked Data[J]. *Information Studies: Theory & Application*, 2012,35(2):46-51.)

[26] Schandl T, Blumauer A. PoolParty: SKOS Thesaurus Management Utilizing Linked Data[C]. In: *Proceedings of the 7th Extended Semantic Web Conference (ESWC2010)*. 2010:421-425.

[27] Clarke C. A Resource List Management Tool for Undergraduate Students Based on Linked Open Data Principles[C]. In: *Proceedings of the 6th European Sematic Web Conference*. 2009: 697-707.

[28] Jessop D M, Adams S E, Murray-Rust P. Mining Chemical Information from Open Patents[J]. *Journal of Cheminformatics*, 2011, 3(1):40.

[29] Pesenhofer A, Berger H, Dittenbach M. Offering New Insights by Harmonizing Patents, Taxonomies and Linked Data[A]. //Current Challenges in Patent Information Retrieval[M]. Springer, 2011:357-371.

[30] Damova M, Dannells D. Reason-able View of Linked Data for Cultural Heritage[C]. In: *Proceedings of the 3rd International Conference on Software, Services and Semantic Technologies*. 2011: 17-24.

[31] Davis C, Nikolic I, Dijkema G P J. Industrial Ecology 2.0[J]. *Journal of Industrial Ecology*, 2010, 14(5):707.

[32] 沈志宏,张晓林. 语义网环境下数据溯源表达模型研究综述[J]. *现代图书情报技术*, 2011 (4):1-8.(Shen Zhihong, Zhang Xiaolin. Data Provenance Model in Semantic Web Environment: An Overview[J]. *New Technology of Library and Information Service*,2011(4):1-8.)

(作者 E-mail:gulp@mail.las.ac.cn)